Press Release

FUTURE

2025年9月2日
フューチャー株式会社
(東証プライム:証券コード 4722)

フューチャー、ソフトウェア開発に関する
世界最大規模の日本語インストラクションチューニングデータを公開
大規模言語モデルとソフトウェア開発の研究に貢献

フューチャー株式会社(本社:東京都品川区、代表取締役会長兼社長 グループCEO 金丸恭文、以下フューチャー)は、大規模言語モデル(以下、LLM)と日本語によるソフトウェア開発領域の研究の発展を目的に、ソフトウェア開発に関するインストラクションチューニング(Instruction-Tuning)データを無償公開しました。公開したのは、シングルターン※1の日本語530万件、英語610万件、マルチターン※1の英語85万件のデータセットで、ソフトウェア開発に関する日本語インストラクションチューニングデータでは世界最大規模です。

◆公開URL◆
・シングルターン: https://huggingface.co/datasets/future-architect/Llama-3.3-Future-Code-Instructions
・マルチターン: https://huggingface.co/datasets/future-architect/Llama-3.3-Future-Code-Instructions-MT

LLMの開発には良質な学習データが不可欠です。特に、人がLLMに与える指示(Instruction)とそれに対する回答(Answer)のペアで構成されるインストラクションチューニングデータは非常に重要です。しかし、通常、これらのデータ構築には多額のコストが掛かるため、一般公開されている学習用データセットは少なく、かつ日本語に特化したソフトウェアに関するインストラクションチューニングデータも限られていることが同分野の研究開発における障害となっています。

当社は、2024年10月に経済産業省とNEDO(国立研究開発法人 新エネルギー・産業技術総合開発機構)が実施する国内生成AIの開発力強化プロジェクト「GENIAC(Generative AI Accelerator Challenge)」※2に採択され「日本語とソフトウェア開発に特化した基盤モデル」の研究開発を行ってきました。今回公開したインストラクションチューニングデータは、本プロジェクトの研究過程においてベンチマークとしたLLMをもとに自動生成したものです。なお本インストラクションチューニングデータを活用し、GENIACのプロジェクトで開発した「Llama 3.1 Future Code Ja」※3は、様々なプログラミング言語において比較対象となるベースモデルに対し高い生成能力を実現し、特に日本語の指示によるソースコード補完能力に優れていることが確認できました。

フューチャーはテクノロジーをベースに、20社以上のグループ企業を展開するソーシャルデザインカンパニーとして、先端的なAIの学術研究・研究開発を強力に推し進めています。また、主要事業会社のフューチャーアーキテクト(本社:東京都品川区、代表取締役社長:谷口友彦)とともに、構想から実装まで一気通貫で支援するAI コンサルティングサービスを提供しています。

公開データの規模感

FUTURE

データセット	日本語インストラクション数
future-architect/Llama-3.3-Future-Code-Instructions	**531.4万**
Aratako/Synthetic-JP-EN-Coding-Dataset-801k	17.3万
HachiML/alpaca_jp_python	1.0万

以下の出典をもとに当社作成
・Aratako/Synthetic-JP-EN-Coding-Dataset-801k https://huggingface.co/datasets/Aratako/Synthetic-JP-EN-Coding-Dataset-801k
・HachiML/alpaca_jp_python https://huggingface.co/datasets/HachiML/alpaca_jp_python

当社は世界最大規模のソフトウェア開発に関する日本語インストラクションチューニングデータを公開することで、日本語のソフトウェア開発の研究と発展に貢献していきます。今後も、「最先端のAI研究の推進」と「AI技術の社会実装」を強みに、AI利活用と社会実装のNo1 カンパニーを目指します。

公開データの例

In Bash, how would I convert a hexadecimal string to a decimal integer?

````
```shell
#!/bin/bash

hex_to_int() {
  echo $((16#$1))
}

# Test the function
echo $(hex_to_int '1a')  # Output: 26
echo $(hex_to_int 'ff')  # Output: 255
```
````

Bashでは、16進数文字列を10進数整数に変換するにはどうすればよいですか？

````
```shell
#!/bin/bash

hex_to_int() {
  echo $((16#$1))
}

# 関数をテストします
echo $(hex_to_int '1a')  # 出力: 26
echo $(hex_to_int 'ff')  # 出力: 255
```
````

※1. シングルターンは質問と回答のペアで、一問一答形式となっているデータ。マルチターンは、 複数の質問応答のやり取りを含むデータ。

※2. GENIAC https://www.meti.go.jp/policy/mono_info_service/geniac/index.html
NEDO採択決定 https://www.nedo.go.jp/koubo/IT3_100331.html
＜参考プレスリリース＞GENIAC 採択「日本語とソフトウェア開発に特化した基盤モデル」を一般公開
https://www.future.co.jp/press_room/PDF/PressRelease_FC_GENIAC_20250609.pdf

※3.「Llama 3.1 Future Code Ja」
https://huggingface.co/future-architect/Llama-3.1-Future-Code-Ja-8B

■本件に関する報道機関からのお問合せ先
フューチャー株式会社　広報担当：清水、石井
TEL：03-5740-5721
お問い合わせフォーム：https://www.future.co.jp/apps/contact/corp/press_interview_entry.php